Interpretation, and Applications of Gradients

Part 3: Gradients as Uncertainty

Objectives

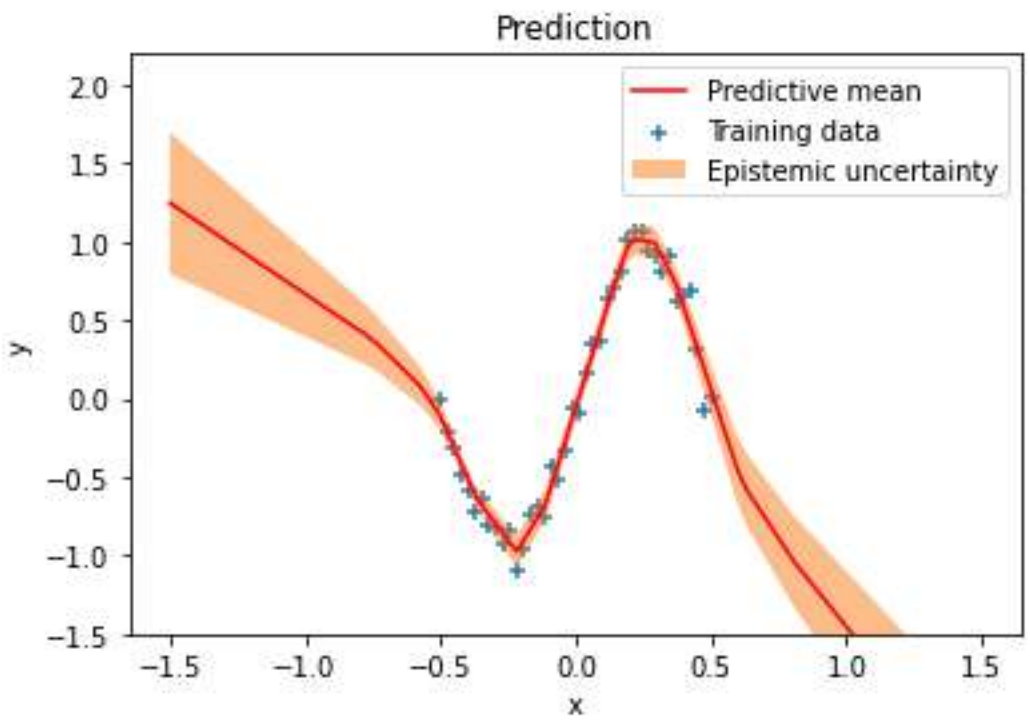
Objectives in Part 3

- Interpret gradients as Uncertainty
- Uncertainty Applications
  - Anomaly Detection
  - Out-of-Distribution Detection
  - Adversarial Image Detection
  - Corruption Detection

Uncertainty

What is Uncertainty?

Uncertainty is a model knowing that it does not know



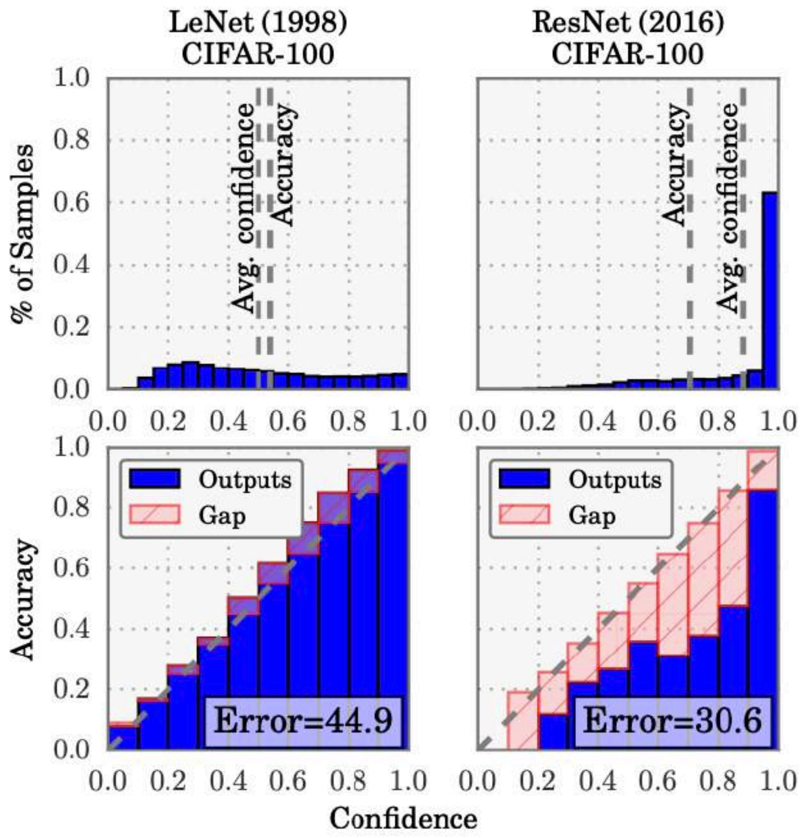
A simple example: More the training data, lesser the uncertainty

http://krasserm.github.io/2020/09/25/reliable-uncertainty-estimates/

Uncertainty

When is Uncertainty an Issue?

Uncertainty is a model knowing that it does not know

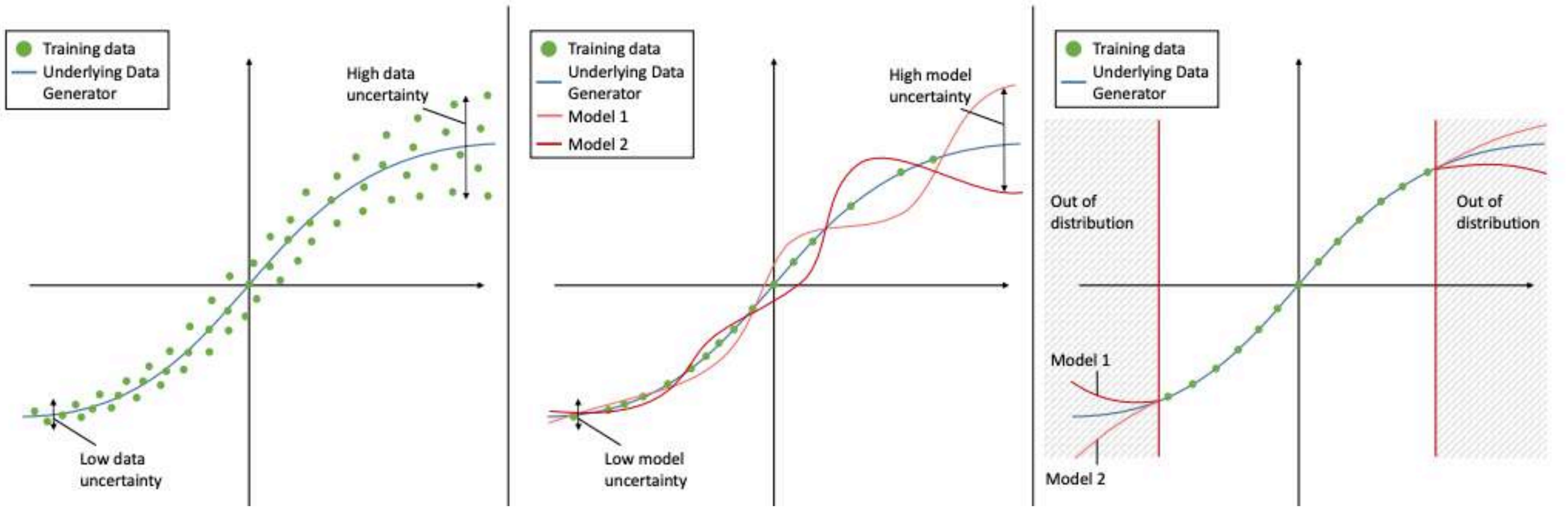


- Larger the model, more misplaced is a network's confidence
- On ResNet, the gap between prediction accuracy and its corresponding confidence is significantly high
- On OOD data, uncertainty is not easy to quantify

Guo, Chuan, et al. "On calibration of modern neural networks." *International conference on machine learning*. PMLR, 2017.

Uncertainty
Two Types of Uncertainty

Two major types of uncertainty: Uncertainty in data and uncertainty in model, together termed as prediction Uncertainty

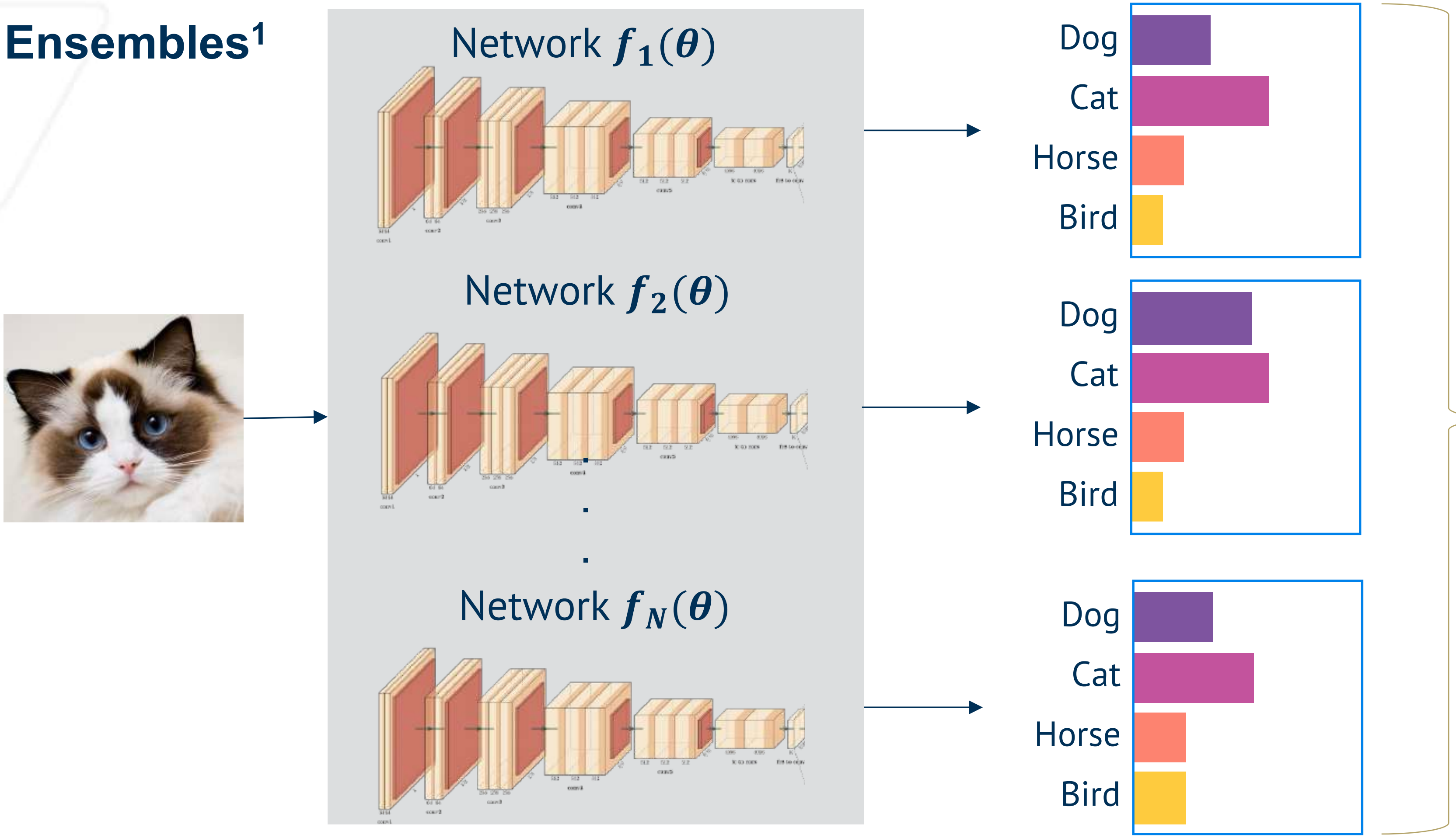


Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., ... & Zhu, X. X. (2021). A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*.

Uncertainty

Uncertainty Quantification in Neural Networks

Via Ensembles[1]



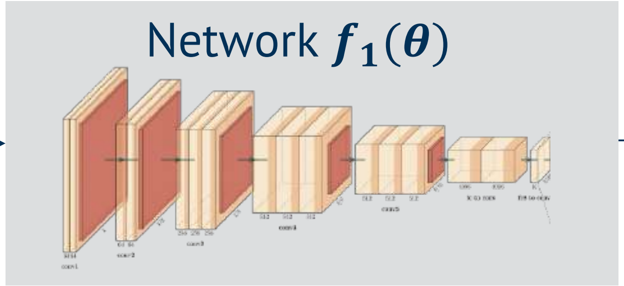
Variation within outputs $Var(y)$ is the uncertainty. Commonly referred to as **Prediction Uncertainty.**

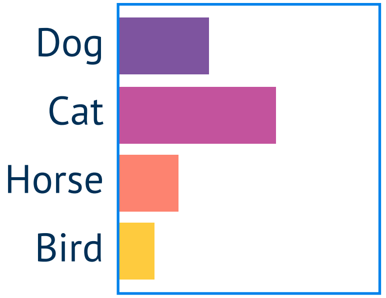
[1] Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles." *Advances in neural information processing systems* 30 (2017).

Uncertainty

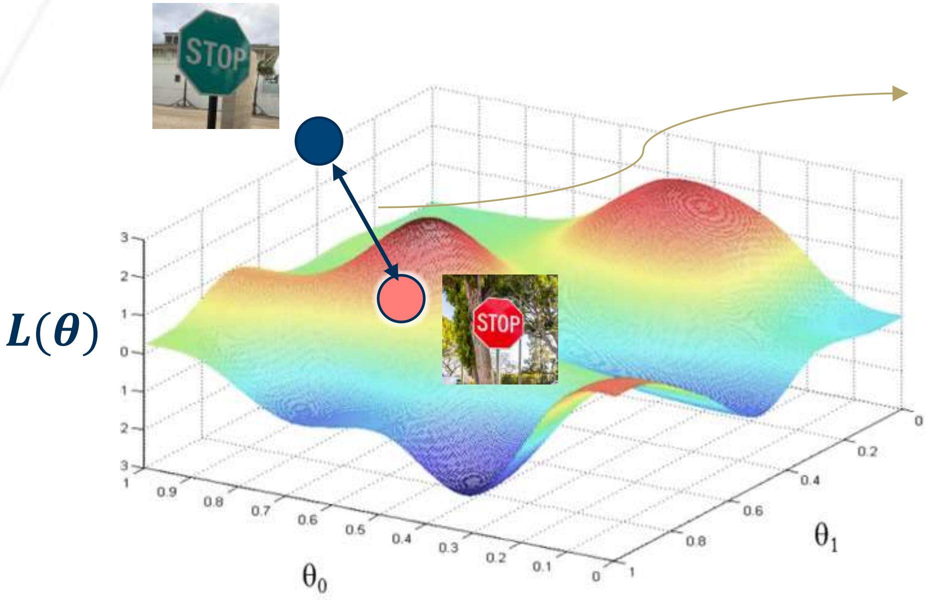
Uncertainty Quantification in Neural Networks

Via Single pass methods[1]





Uncertainty quantification using a single network and a single pass



Calculate distance from some trained clusters

Does not require multiple networks!
However, does requires multiple data points at inference!

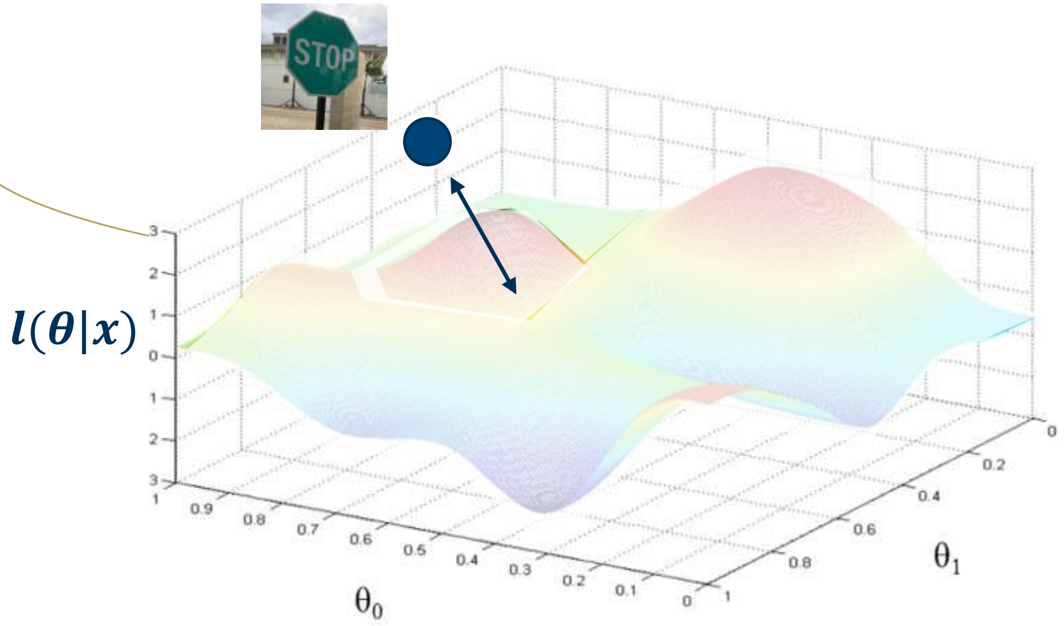
[1Van Amersfoort, J., Smith, L., Teh, Y. W., & Gal, Y. (2020, November). Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning* (pp. 9690-9700). PMLR.

Uncertainty

Gradients as Single pass Features

Our Goal: Use gradients to characterize the novel data at Inference, without global information

Distance from unknown cluster



Two techniques:

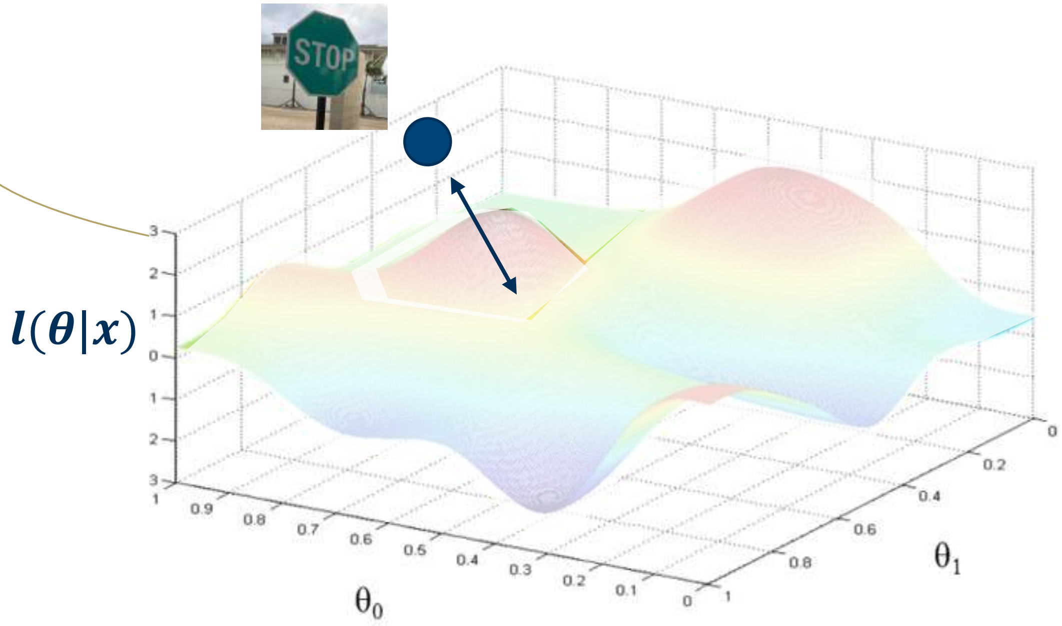
1. Gradient constraints during Training for Anomaly Detection
2. Backpropagating Confounding labels for Out-of-Distribution Detection

Uncertainty

Gradients as Single pass Features

Our Goal: Use gradients to characterize the novel data at Inference, without global information

Distance from unknown cluster



Two techniques:

1. Gradient constraints during Training for Anomaly Detection
2. Backpropagating Confounding labels for Out-of-Distribution Detection

Backpropagated Gradient Representations for Anomaly Detection

Gukyeong Kwon, PhD
Amazon AWS

Mohit Prabhushankar, PhD
Postdoc, Georgia Tech

Ghassan AlRegib, PhD
Professor, Georgia Tech

SCAN ME

Anomalies

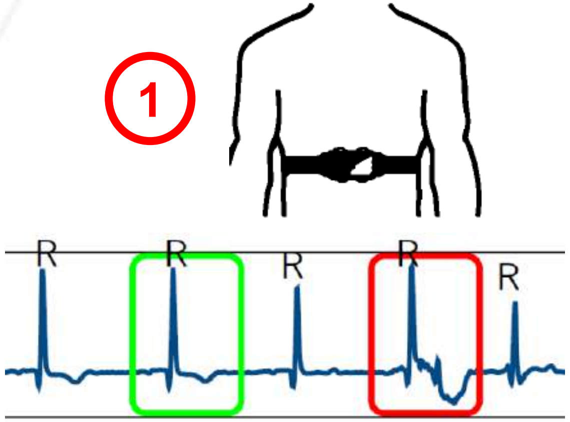
Finding Rare Events in Normal Patterns

Backpropagated Gradient Representations for Anomaly Detection

SCAN ME

'Anomalies are patterns in data that do not conform to a well defined notion of normal behavior' [1]

1



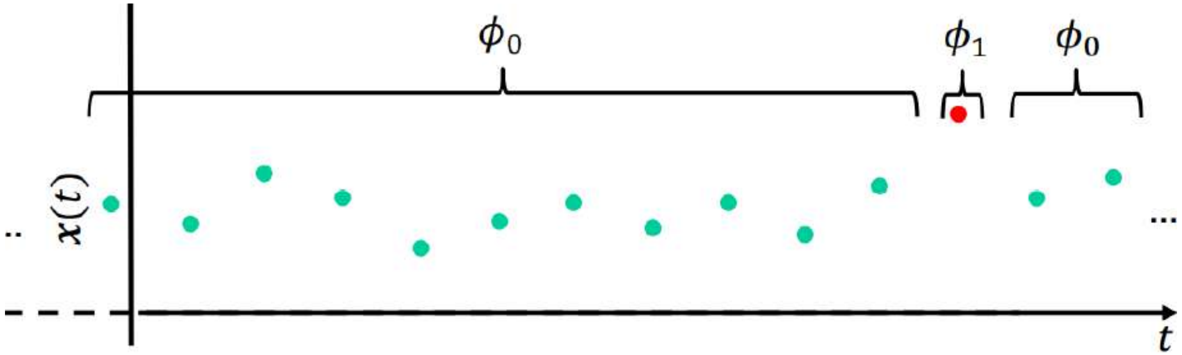
2

Statistical Definition:

- Normal data are generated from a stationary process P_N
- Anomalies are generated from a different process $P_A \neq P_N$

Goal: Detect ϕ_1

$$x(t) = \begin{cases} \phi_0 & \text{Normal data} \\ \phi_1 & \text{Anomalies} \end{cases}$$



[1] V. Chandola, A. Banerjee, V. Kumar. "Anomaly detection: A survey". ACM Comput. Surv. 41, 3, Article 15 (July 2009), 58 pages

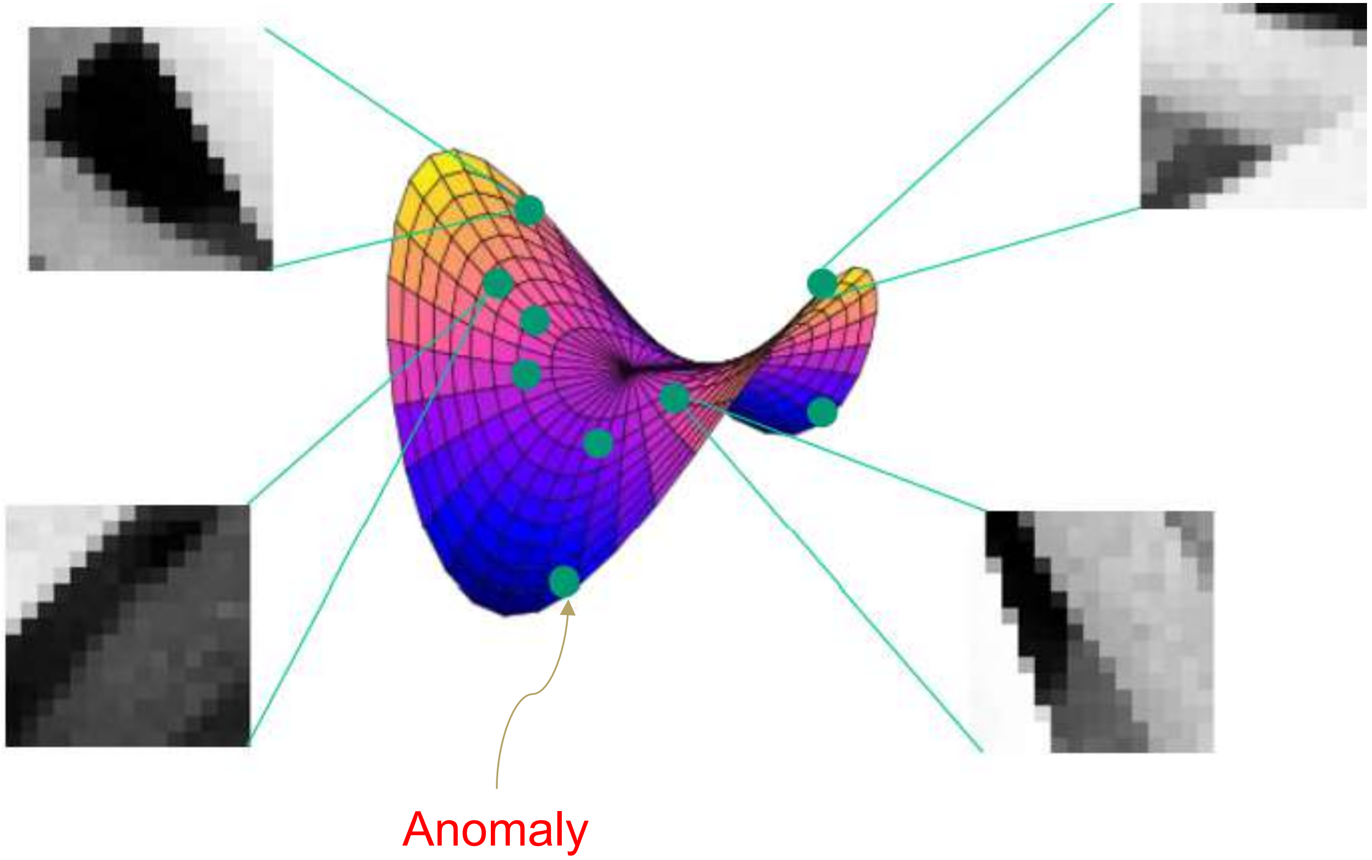
Anomalies
Steps for Anomaly Detection

Backpropagated Gradient Representations for Anomaly Detection

SCAN ME

Step 1: Constrain manifolds, Step 2: Detect statistically implausible projections

- Step 1 ensures that patches from natural images live close to a low dimensional manifold
- Step 2 designs distance functions that detect *implausibility* based on constraints



Constraining Manifolds

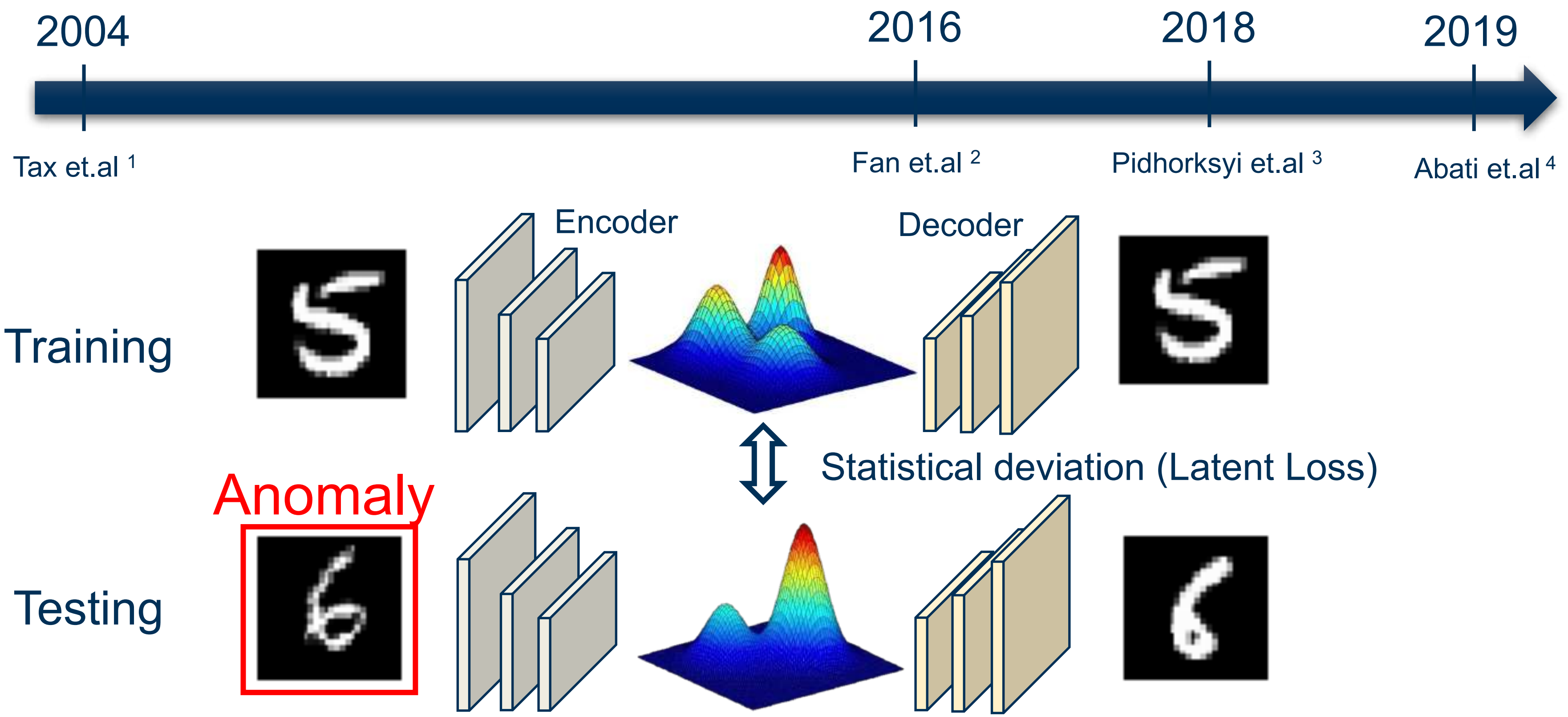
General Constraints

Backpropagated Gradient Representations for Anomaly Detection

SCAN ME

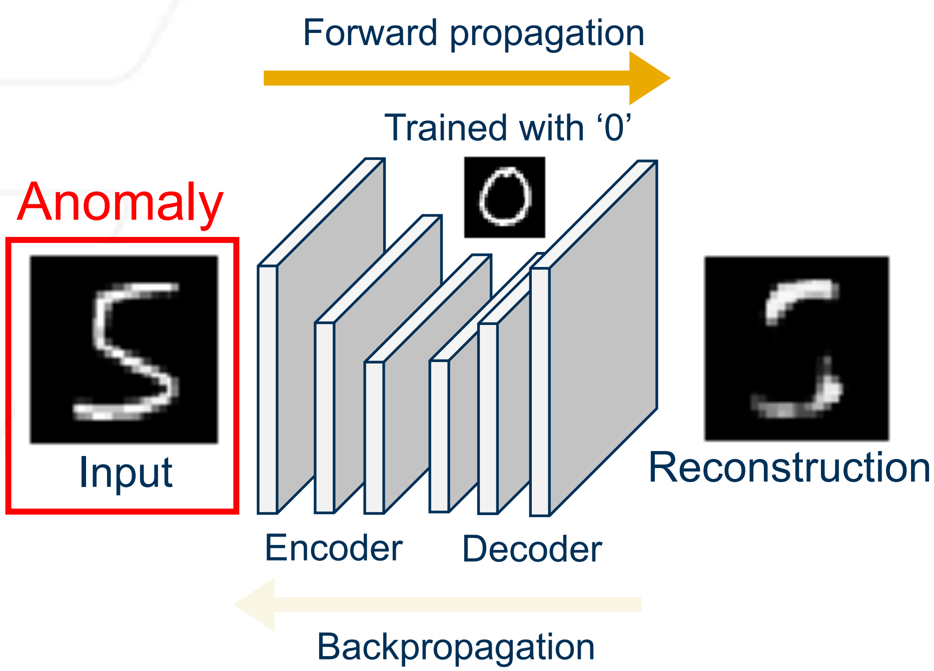
Constrained Representation

Activations are constrained using GANs, VAEs, etc.



[1] David MJ Tax and Robert PW Duin. Support vector data description. Machine learning, 54(1):45–66, 2004.
[2] Yaxiang Fan, Gongjian Wen, Deren Li, Shaohua Qiu, and Martin D Levine. Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder. arXiv preprint arXiv:1805.11223, 2018. 1, 2
[3] S. Pidhorskyi, R. Almohsen, and G. Doretto, “Generative probabilistic novelty detection with adversarial autoencoders,” in Advances in Neural Information Processing Systems, 2018, pp. 6822–6833.
[4] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, “Latent space autoregression for novelty detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 481–490.

Constraining Manifolds
Gradient-based Constraints



Activation Constraints

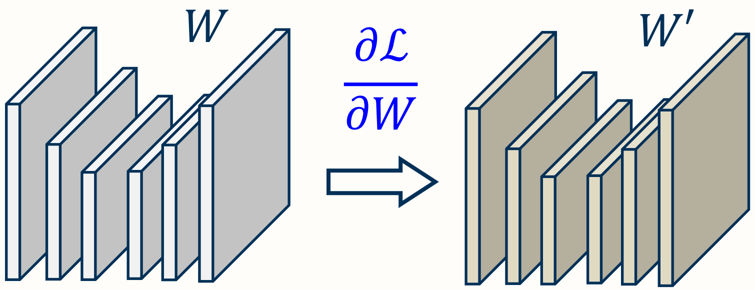
Activation-based representation
(Data perspective)

e.g. Reconstruction error ($\mathcal{L}$)

How much of the input does not correspond to the learned information?

Gradient Constraints

Gradient-based Representation
(**Model** perspective)



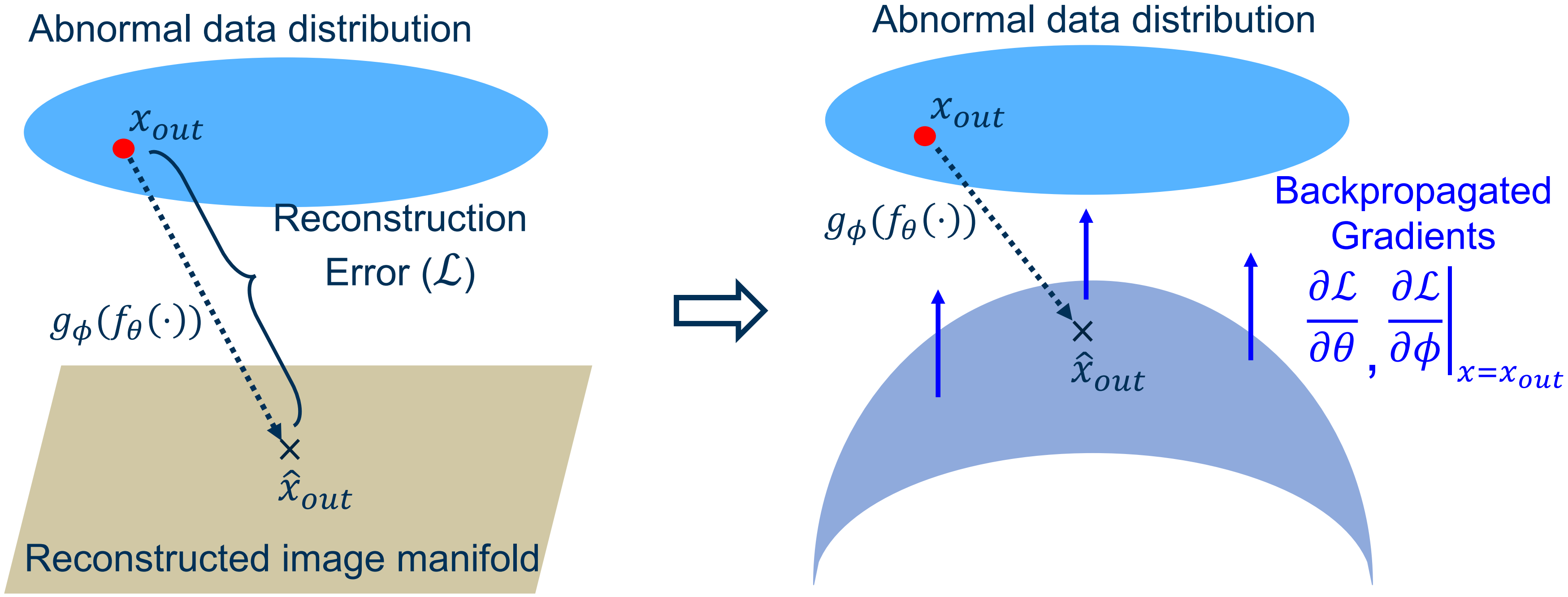
How much **model update** is required by the input?

G. Kwon, M. Prabhushankar, D. Temel, and G. AlRegib, "Backpropagated Gradient Representations for Anomaly Detection," 2020

Constraining Manifolds

Advantages of Gradient-based Constraints

- **Gradients provide directional information to characterize anomalies**
- **Gradients from different layers capture abnormality at different levels of data abstraction**



G. Kwon, M. Prabhushankar, D. Temel, and G. AlRegib, "Backpropagated Gradient Representations for Anomaly Detection," 2020

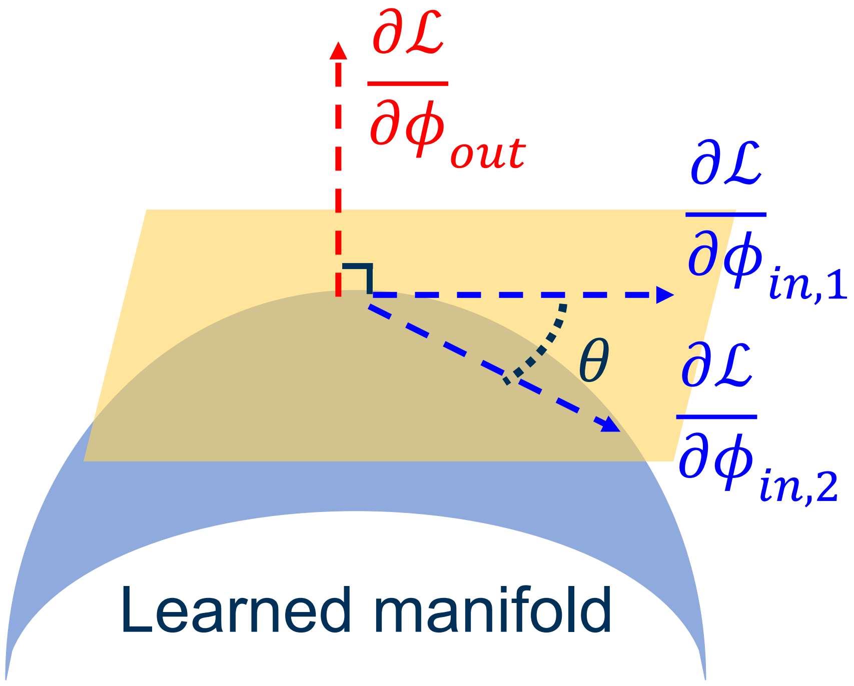
GradCON: Gradient Constraint

Gradient-based Constraints

Backpropagated Gradient Representations for Anomaly Detection

SCAN ME

Constrain gradient-based representations during training to obtain clear separation between normal data and abnormal data



ϕ: Weights $\mathcal{L}$: Reconstruction error

At k-th step of training,

$$J = \mathcal{L} - \mathbb{E}_i\left[\underbrace{\text{cosSIM}\left(\frac{\partial J}{\partial \phi_{i_{avg}}}^{k-1}, \frac{\partial \mathcal{L}}{\partial \phi_i}^{k}\right)}_{\text{Gradient loss}}\right]$$

Avg. training gradients until (k-1) th iter.

Gradients at k-th iter.

where $$\frac{\partial J}{\partial \phi_{i_{avg}}}^{k-1} = \sum_{t=1}^{k-1} \frac{\partial J}{\partial \phi_i}^{t}$$

G. Kwon, M. Prabhushankar, D. Temel, and G. AlRegib, "Backpropagated Gradient Representations for Anomaly Detection," 2020

GradCON: Gradient Constraint

Activations vs Gradients

SCAN ME

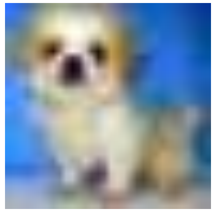
Backpropagated Gradient Representations for Anomaly Detection

AUROC Results

Abnormal “class” detection (CIFAR-10)

e.g.

Normal



Abnormal

Model	Loss	Plane	Car	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck	Average
CAE	Recon	0.682	0.353	0.638	0.587	0.669	**0.613**	0.495	0.498	0.711	0.390	0.564
CAE + Grad	Recon	0.659	0.356	**0.640**	0.555	0.695	0.554	0.549	0.478	0.695	0.357	0.554
	Grad	**0.752**	0.619	0.622	0.580	0.705	0.591	0.683	**0.576**	**0.774**	**0.709**	**0.661**
VAE	Recon	0.553	0.608	0.437	0.546	0.393	0.531	0.489	0.515	0.552	0.631	0.526
	Latent	0.634	0.442	**0.640**	0.497	**0.743**	0.515	**0.745**	0.527	0.674	0.416	0.583
VAE + Grad	Recon	0.556	0.606	0.438	0.548	0.392	0.543	0.496	0.518	0.552	0.631	0.528
	Latent	0.586	0.396	0.618	0.476	0.719	0.474	0.698	0.537	0.586	0.413	0.550
	Grad	0.736	**0.625**	0.591	**0.596**	0.707	0.570	0.740	0.543	0.738	0.629	0.647

Recon: Reconstruction error, Latent: Latent loss, Grad: Gradient loss

- (CAE vs. CAE + Grad) Effectiveness of the gradient constraint
- (CAE vs. VAE) Performance sacrifice from the latent constraint
- (VAE vs. VAE + Grad) Complementary features from the gradient constraint

G. Kwon, M. Prabhushankar, D. Temel, and G. AlRegib, "Backpropagated Gradient Representations for Anomaly Detection," 2020

GradCON: Gradient Constraint

Aberrant Condition Detection

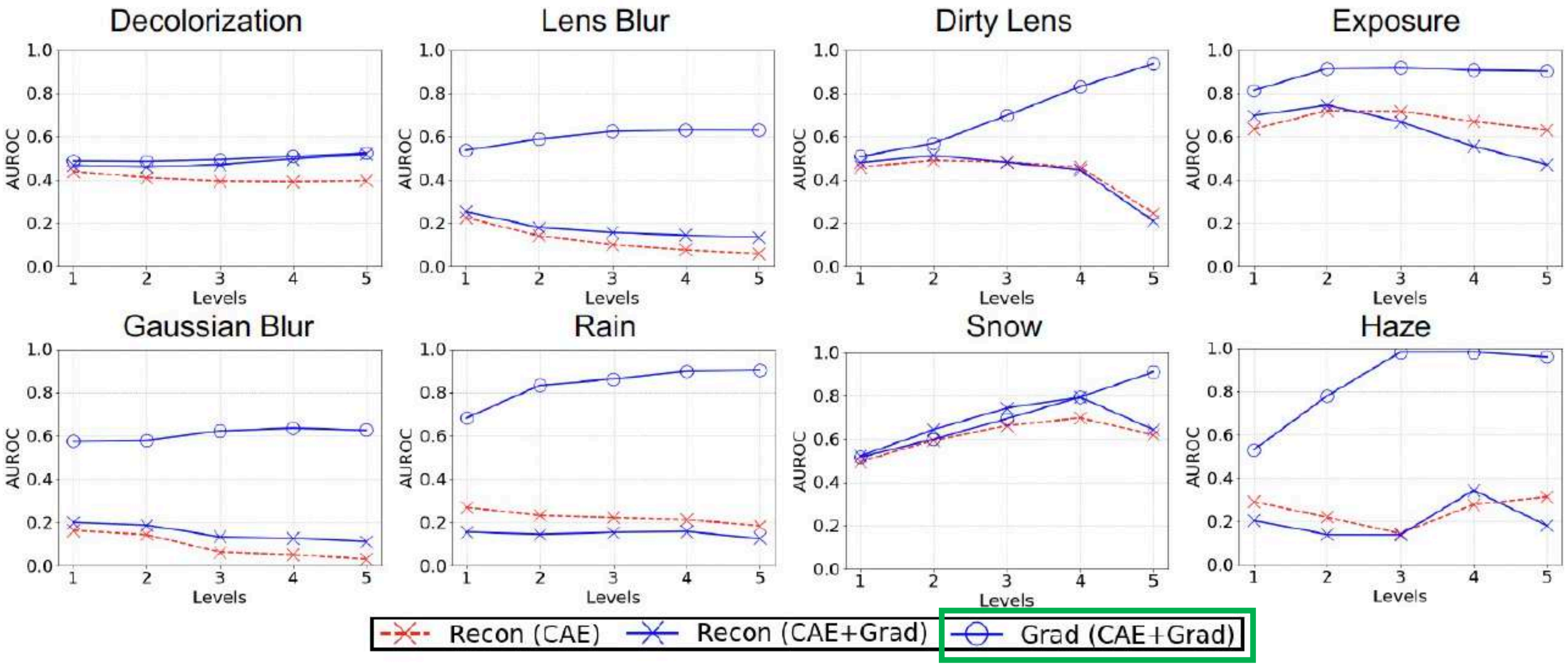
Backpropagated Gradient Representations for Anomaly Detection

SCAN ME

Abnormal “condition” detection (CURE-TSR)

Normal Abnormal

AUROC Results



Recon: Reconstruction error, Grad: Gradient loss

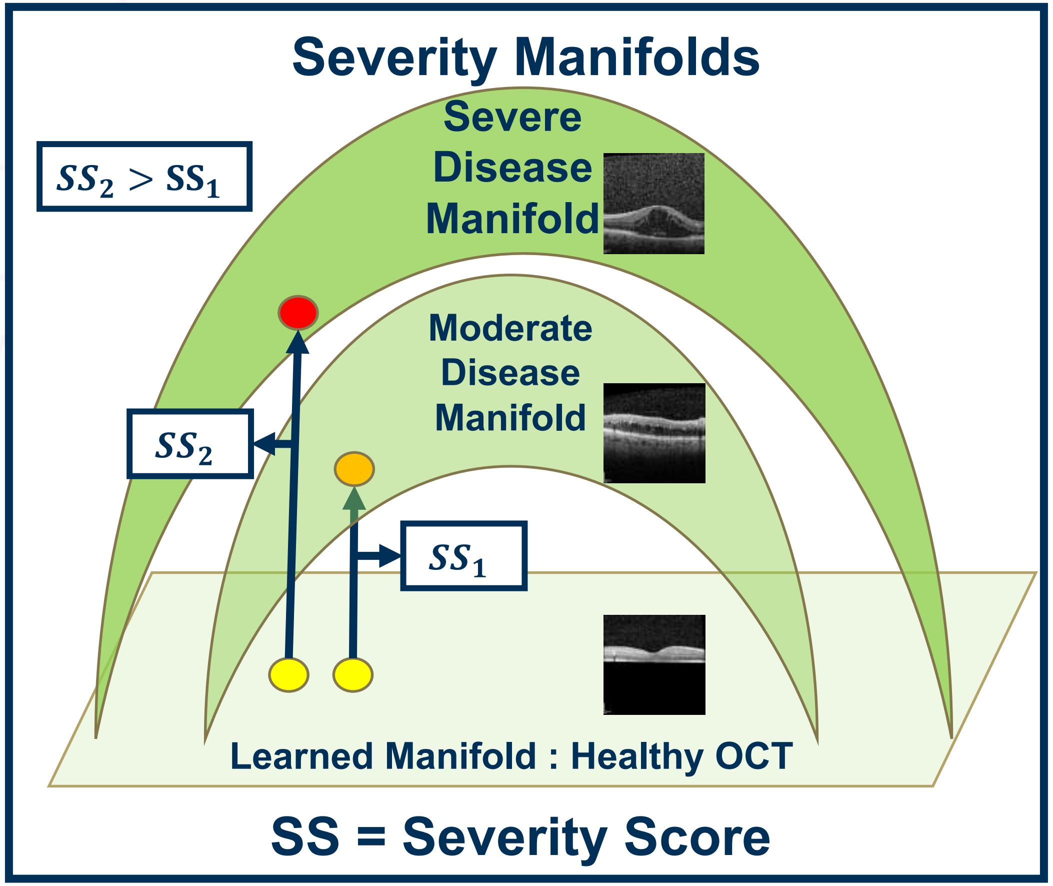
G. Kwon, M. Prabhushankar, D. Temel, and G. AlRegib, "Backpropagated Gradient Representations for Anomaly Detection," 2020

GradCON Applicability

Estimating Disease Severity

Backpropagated Gradient Representations for Anomaly Detection

SCAN ME



Goal

- Define severity with respect to distance from a healthy manifold.
- This distance can be regarded as a severity score.

How to measure severity score?

- Define severity as: “the degree to which a sample appears anomalous relative to the distribution of healthy images.”

Experimental Plan

- Investigate model responses that can act as good surrogate for severity score

K. Kokilepersaud, M. Prabhushankar, G. AlRegib, S. Trejo Corona, C. Wykoff, "Gradient Based Labeling for Biomarker Classification in OCT," in *IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, Oct. 16-19 2022

Dataset: Ophthalmic Labels for Investigating Visual Eye Semantics

- **9408** images **labeled** with complete biomarker data
- Every image associated with vector indicating presence/absence of **16 potential biomarkers**
- 5 biomarkers exist with sufficient balanced quantities
  - Develop 5 biomarker test sets (PAVF, FAVF, IRF, DME, and IRHRF)

https://github.com/olivesgatech

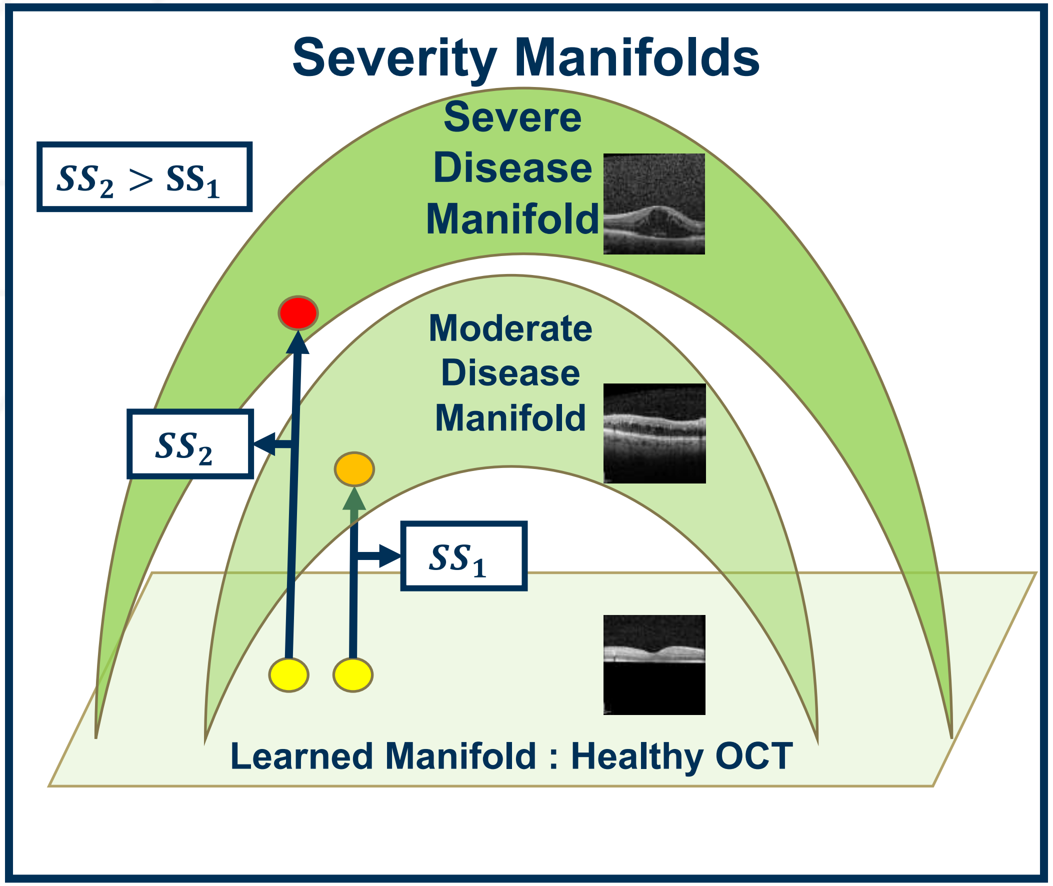
OLIVES Dataset
https://arxiv.org/pdf/2209.11195.pdf

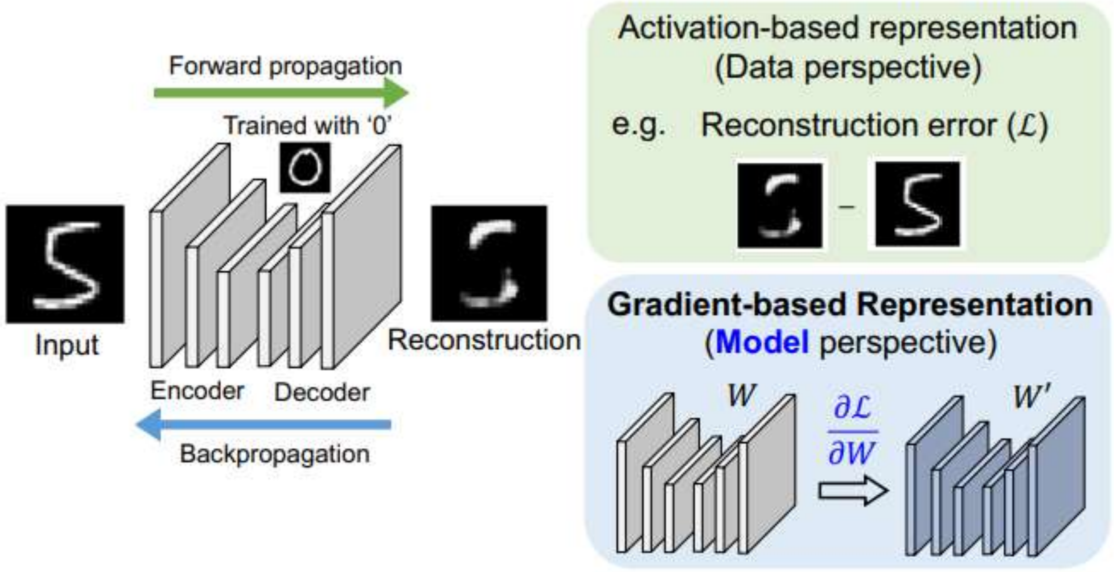
K. Kokilepersaud, M. Prabhushankar, G. AlRegib, S. Trejo Corona, C. Wykoff, "Gradient Based Labeling for Biomarker Classification in OCT," in *IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, Oct. 16-19 2022

GradCON Applicability

Estimating Disease Severity

Backpropagated Gradient Representations for Anomaly Detection





$$\mathcal{L}_{grad} = -\mathop{\mathbb{E}}_{i}\left[\mathrm{cosSIM}\left(\frac{\partial \mathcal{J}}{\partial \phi_i}^{k-1}_{avg}, \frac{\partial \mathcal{L}}{\partial \phi_i}^{k}\right)\right], \quad \frac{\partial \mathcal{J}}{\partial \phi_i}^{k-1}_{avg} = \frac{1}{(k-1)}\sum_{t=1}^{k-1}\frac{\partial \mathcal{J}}{\partial \phi_i}^{t},$$

$$L = L_{recon} + \alpha L_{grad}$$

Idea

- Constrain gradients of in-distribution class
- Make gradients sensitive to progressively anomalous data

K. Kokilepersaud, M. Prabhushankar, G. AlRegib, S. Trejo Corona, C. Wykoff, "Gradient Based Labeling for Biomarker Classification in OCT," in *IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, Oct. 16-19 2022

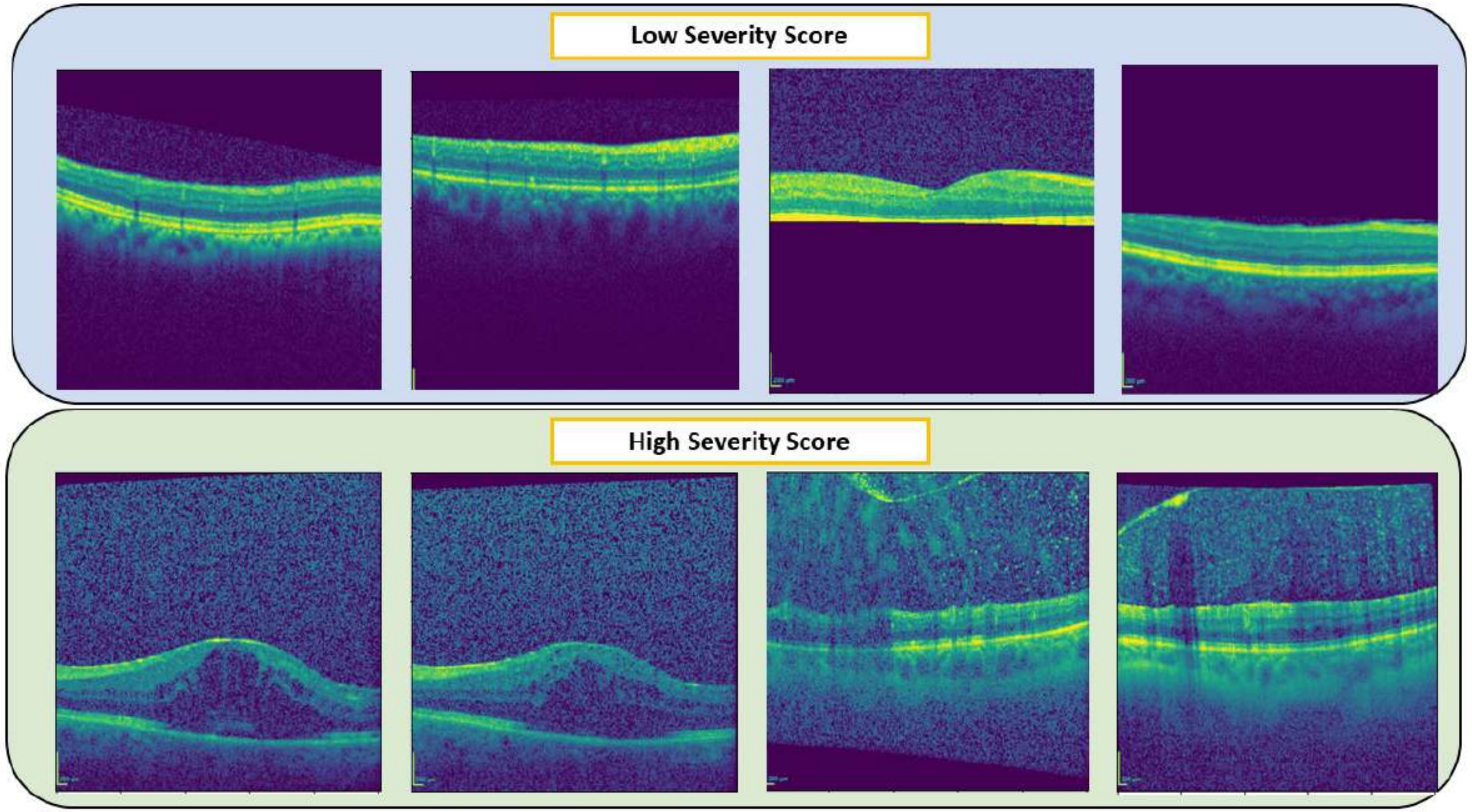
GradCON Applicability

Estimating Disease Severity

Backpropagated Gradient Representations for Anomaly Detection

SCAN ME

Severity Labels used to select positive and negative pairs for weakly-supervised contrastive learning



K. Kokilepersaud, M. Prabhushankar, G. AlRegib, S. Trejo Corona, C. Wykoff, "Gradient Based Labeling for Biomarker Classification in OCT," in *IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, Oct. 16-19 2022

Uncertainty

Gradients as Single pass Features

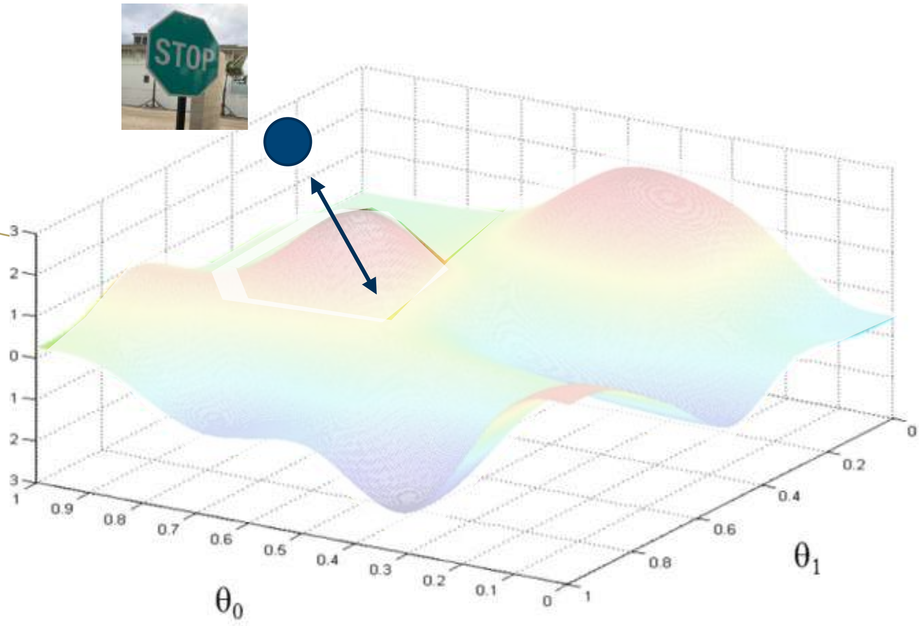
Our Goal: Use gradients to characterize the novel data at Inference, without global information

Distance from unknown cluster

Two techniques:

1. Gradient constraints during Training for Anomaly Detection
2. Backpropagating Confounding labels for Out-of-Distribution Detection

$l(\theta|x)$



Probing the Purview of Neural Networks via Gradient Analysis

Jinsol Lee,
PhD Candidate

Mohit Prabhushankar, PhD
Postdoc

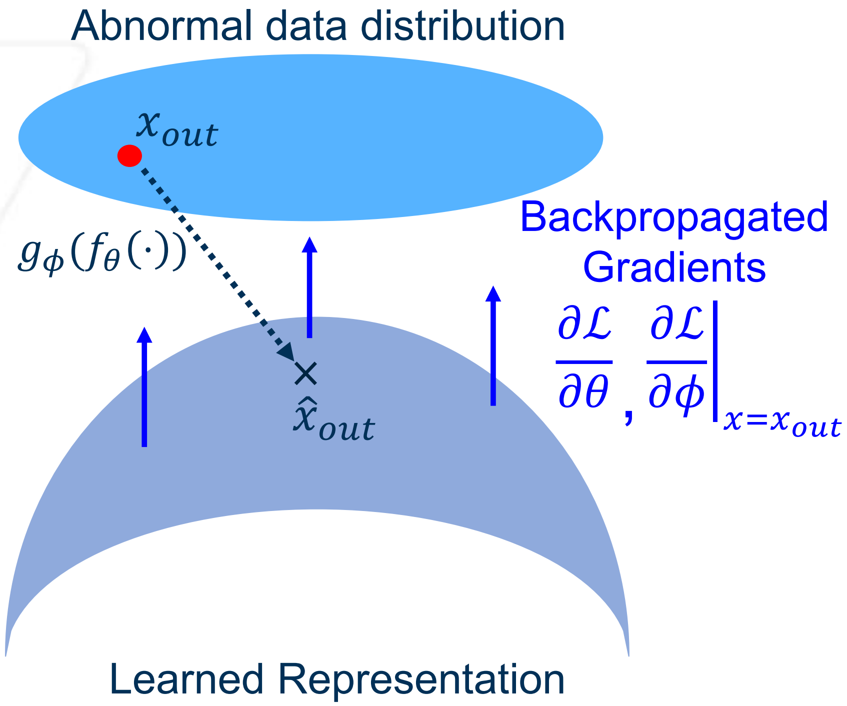
Ghassan AlRegib, PhD
Professor

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

Principle: Gradients provide a distance measure between the learned representations space and novel data



However, what is $\mathcal{L}$?

- In anomaly detection, the loss was between the input and its reconstruction
- In prediction tasks, there is neither the reconstructed input nor ground truth

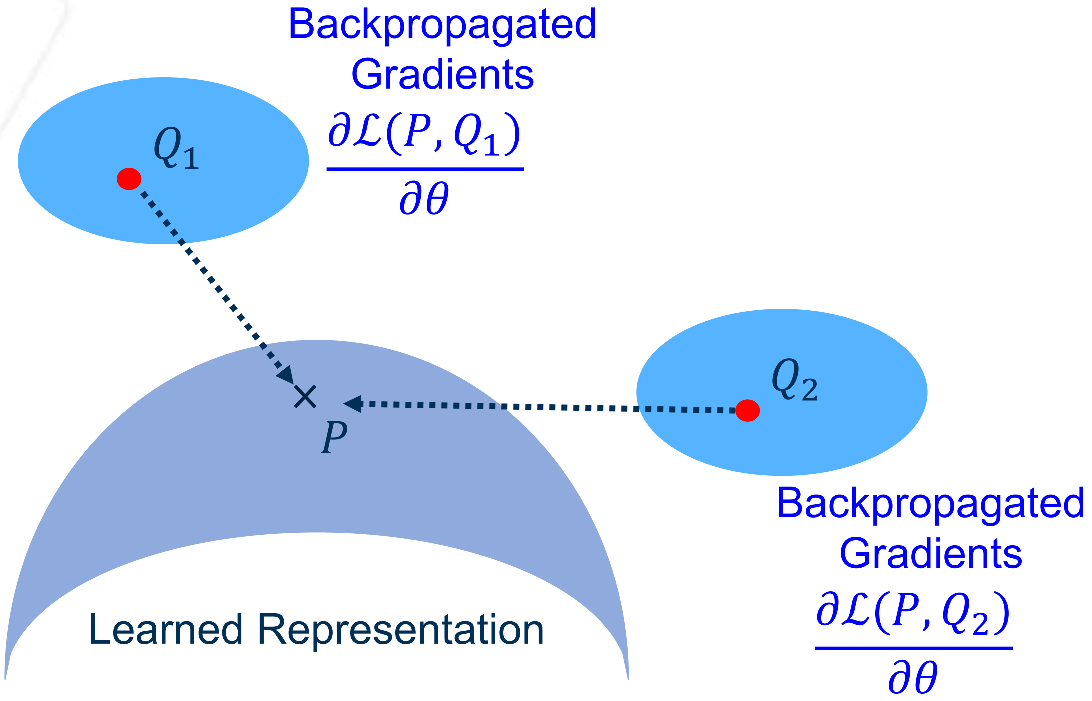
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

Principle: Gradients provide a distance measure between the learned representations space and novel data

$P \;=$ Predicted class
$Q_1 =$ Contrast class 1
$Q_2 =$ Contrast class 2



However, what is $\mathcal{L}$?

- In anomaly detection, the loss was between the input and its reconstruction
- In prediction tasks, there is neither the reconstructed input nor ground truth

- **We backpropagate all contrast classes - $Q_1, Q_2 \ldots Q_N$ by backpropagating N one-hot vectors**
- Higher the distance, higher the uncertainty score

Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

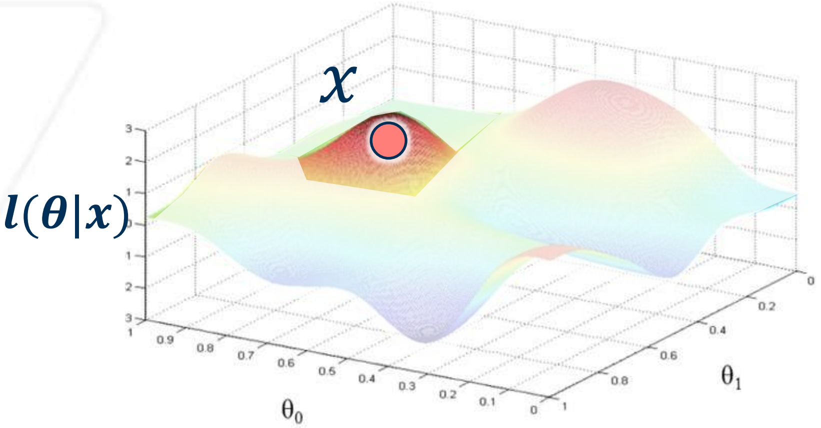
Toy Manifold Example

What is uncertainty?

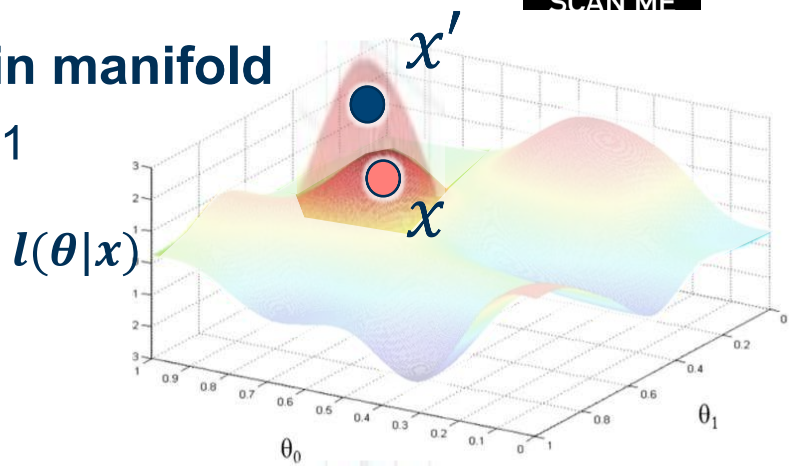
Probing the Purview of Neural Networks via Gradient Analysis

Gradients represent the local required change in manifold

Similar to introspective learning!

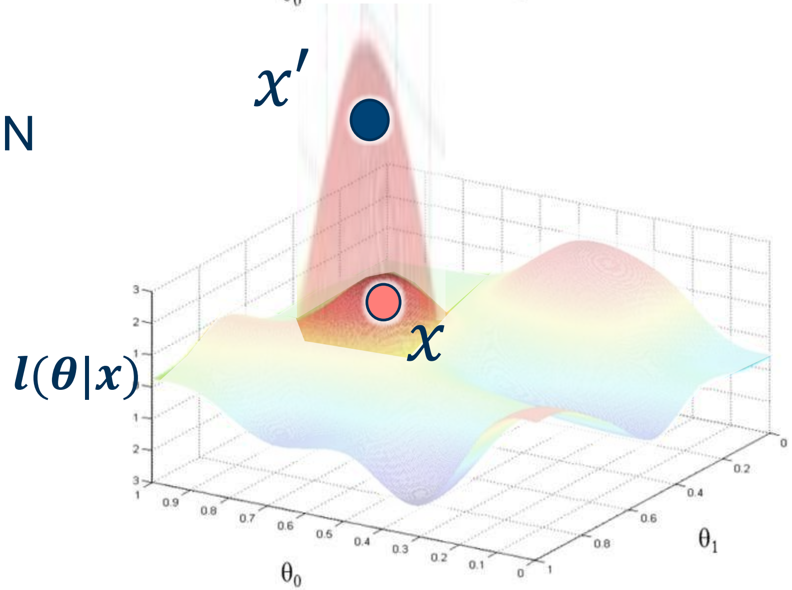


Contrast class 1



.
.
.

Contrast class N



- Gradients provide the necessary change in manifold that would predict the novel data 'correctly'.
- Correctly means contrastively (or incorrectly)!

Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Toy Manifold Example

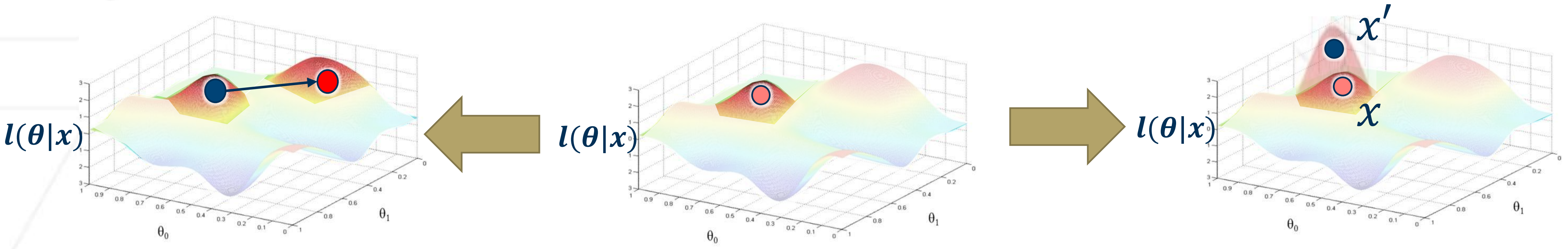
How is this different from Part 2?

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

Part 2: Information

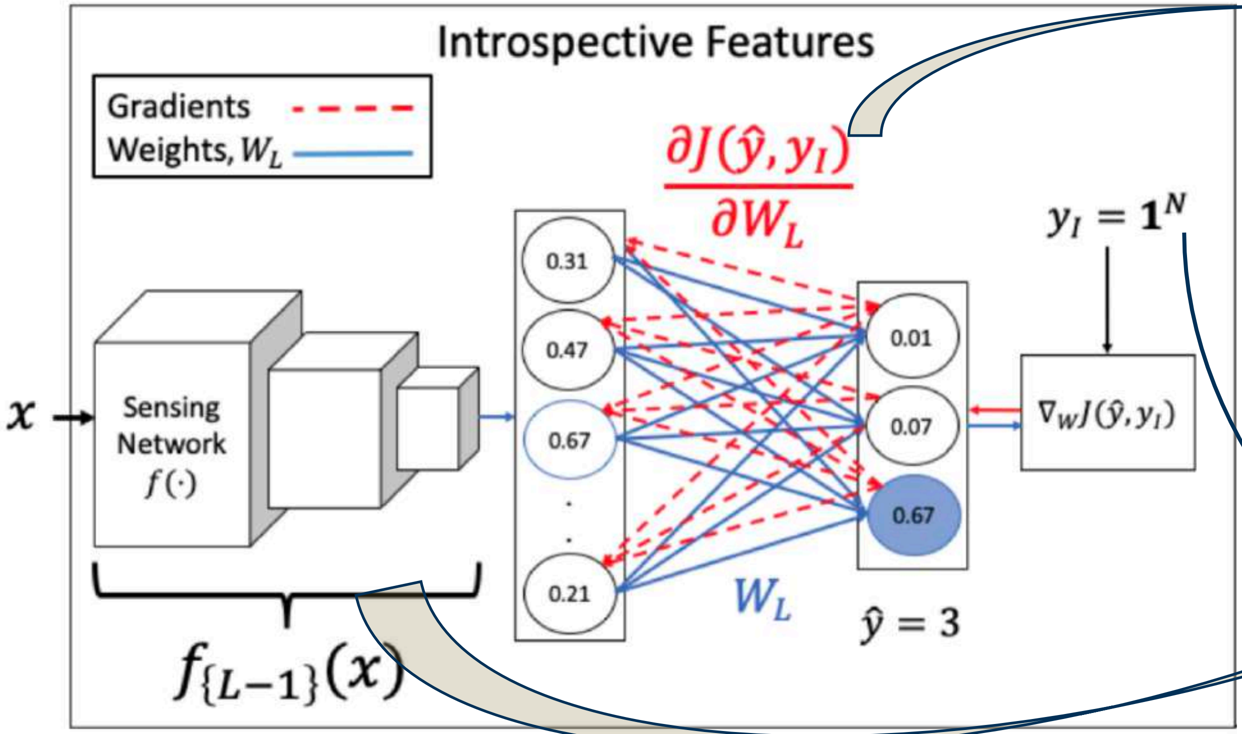
Part 3: Uncertainty



- In Part 2: Activations of learned manifold are weighted by gradients w.r.t. activations to extract information and provide explanations

- In Part 3: Statistics of gradients w.r.t. the weights (energy) will be directly used as features

Step 1: Measure the loss between the prediction P and a vector of all ones and backpropagate to obtain the introspective features



Normalized and vectorized gradients are introspective features.

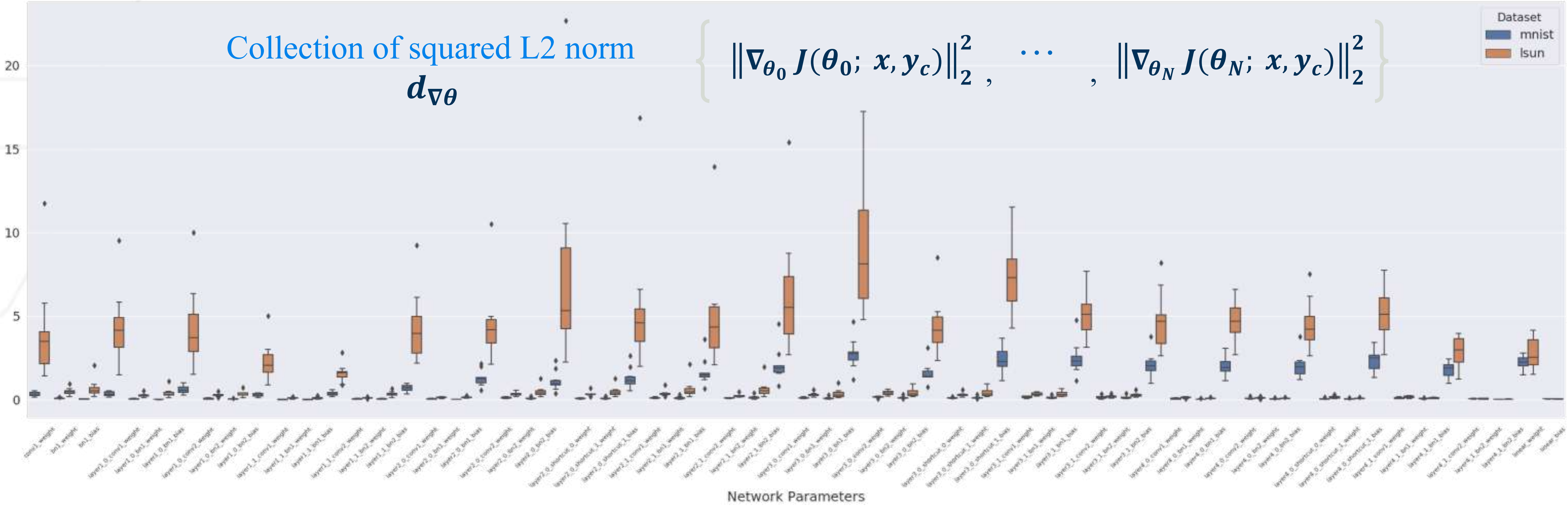
Why vector of all 1s? The theory is presented in [1]

[1] M. Prabhushankar, and G. AlRegib, "Introspective Learning : A Two-Stage Approach for Inference in Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Nov. 29 - Dec. 1 2022.

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

Step 2: Take L2 norm of all generated gradients



MNIST: In-distribution, SUN: Out-of-Distribution

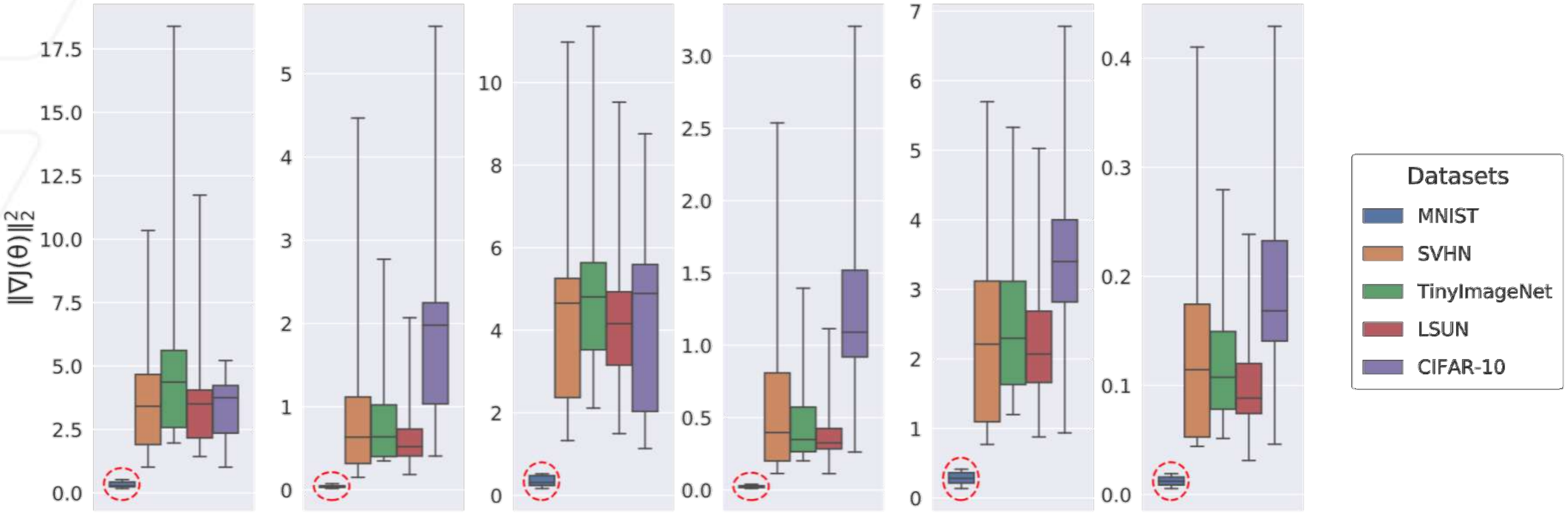
Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Gradient-based Uncertainty

Uncertainty in OOD Setting

SCAN ME

Squared L2 distances for different parameter sets



MNIST: Circled in red. Significantly lower uncertainty compared to OOD datasets

Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

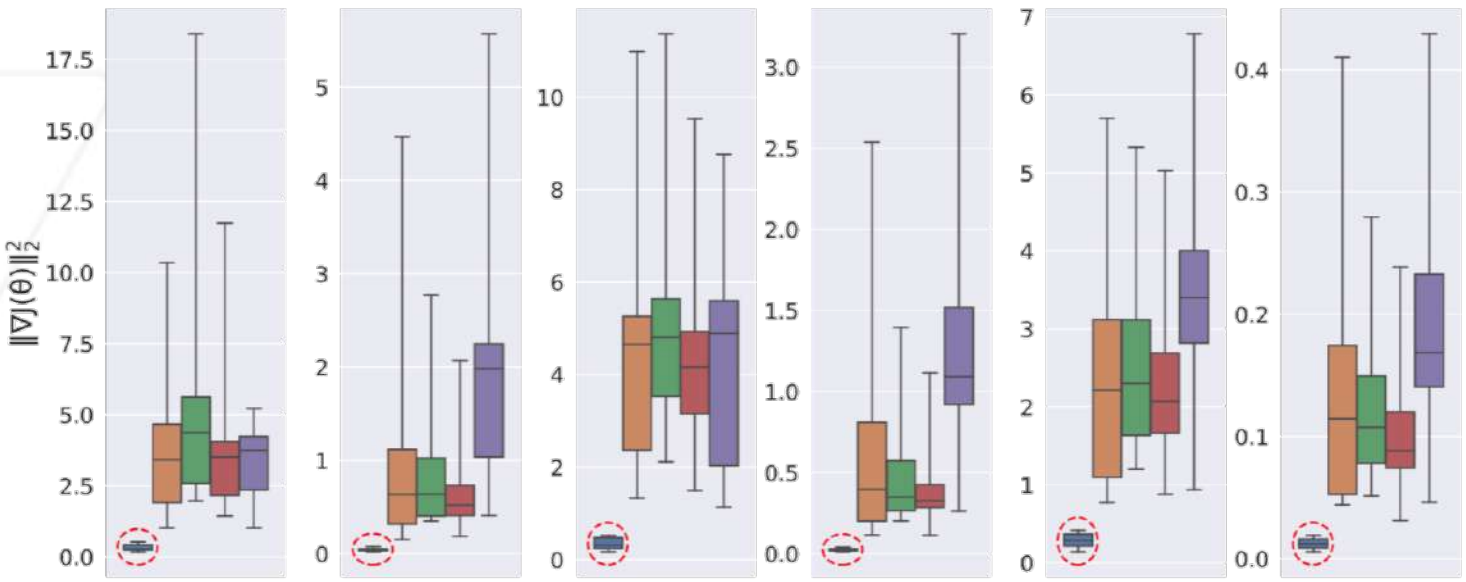
Gradient-based Uncertainty

Experimental Setup

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

Utilize this discrepancy in trained vs untrained data gradient L2 distance to detect adversarial, noisy, and OOD data



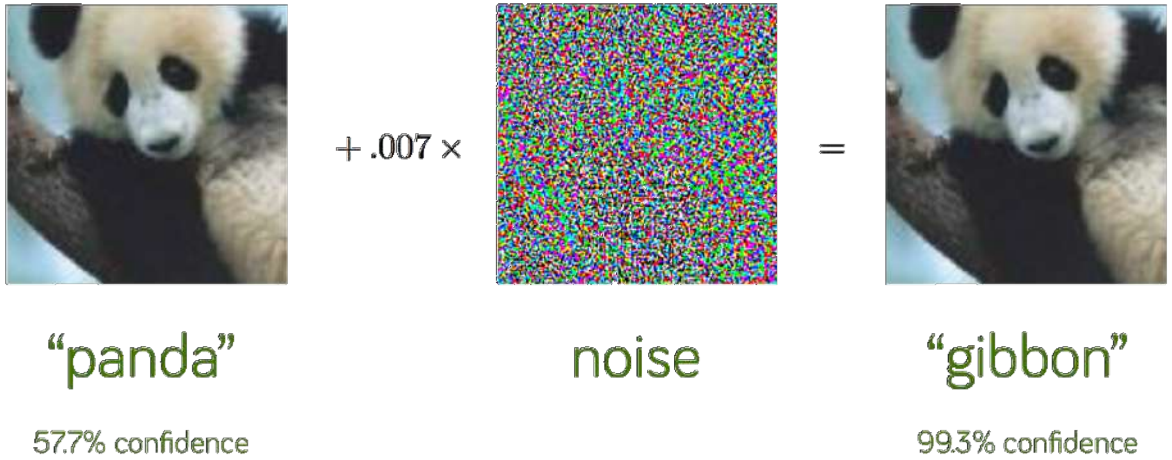
Step 1: Train a deep network $f(\cdot)$ on some **training distribution**
Step 2: Introduce challenging (adversarial, noisy, OOD) data
Step 3: Derive **gradient uncertainty** on both trained and challenge data
Step 4: Train a classifier $H(\cdot)$ to **detect** challenging from trained data
Step 5: At test time, data is passed through $f(\cdot)$ and then $H(\cdot)$ to obtain a **Reliability classification**

Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Gradient-based Uncertainty
Uncertainty in Adversarial Setting

Probing the Purview of Neural Networks via Gradient Analysis

Vulnerable DNNs in the real world



Goal: to examine the ability of trained DNNs to handle adversarial inputs during inference

Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Gradient-based Uncertainty

Uncertainty in Adversarial Setting

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

Model	Attacks	Baseline	LID	M(V)	M(P)	M(FE)	M(P+FE)	Ours
ResNet	FGSM	51.20	90.06	81.69	84.25	**99.95**	**99.95**	93.45
	BIM	49.94	99.21	87.09	89.20	**100.0**	**100.0**	96.19
	C&W	53.40	76.47	74.51	75.71	92.78	92.79	**97.07**
	PGD	50.03	67.48	56.27	57.57	65.23	75.98	**95.82**
	IterLL	60.40	85.17	62.32	64.10	85.10	92.10	**98.17**
	Semantic	52.29	86.25	64.18	65.79	83.95	84.38	**90.15**
DenseNet	FGSM	52.76	98.23	86.88	87.24	**99.98**	99.97	96.83
	BIM	49.67	**100.0**	89.19	89.17	**100.0**	**100.0**	96.85
	C&W	54.53	80.58	75.77	76.16	90.83	90.76	**97.05**
	PGD	49.87	83.01	70.39	66.52	86.94	83.61	**96.77**
	IterLL	55.43	83.16	70.17	66.61	83.20	77.84	**98.53**
	Semantic	53.54	81.41	62.16	62.15	67.98	67.29	**89.55**

Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Gradient-based Uncertainty

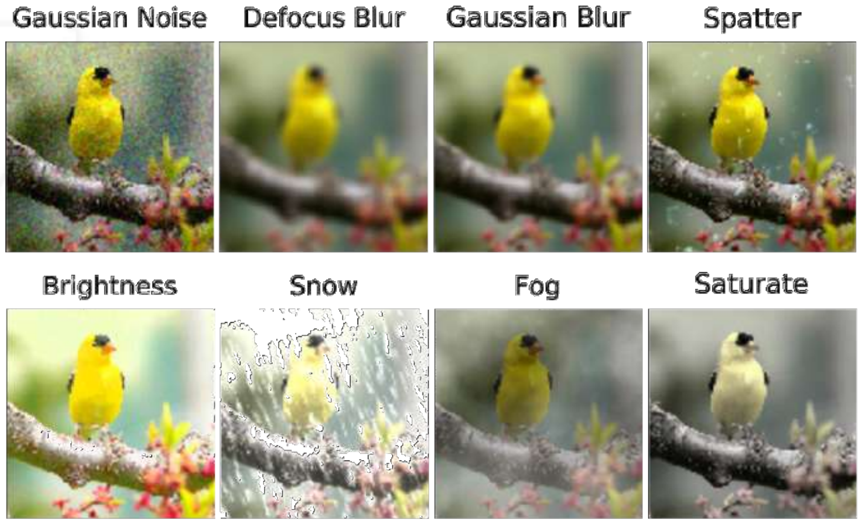
Uncertainty in Detecting Challenging Conditions

SCAN ME

Probing the Purview of Neural Networks via Gradient Analysis

Same application as Anomaly Detection, except there is no need for an additional AE network!

CIFAR-10-C



CURE-TSR

Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Gradient-based Uncertainty

Uncertainty in Detecting Challenging Conditions

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

Dataset	Method / Corruption	Mahalanobis [12] / Ours Level 1	Level 2	Level 3	Level 4	Level 5
CIFAR-10-C	Noise	96.63 / **99.95**	98.73 / **99.97**	99.46 / **99.99**	99.62 / **99.97**	99.71 / **99.99**
	LensBlur	94.22 / **99.95**	97.51 / **99.99**	99.26 / **100.0**	99.78 / **100.0**	99.89 / **100.0**
	GaussianBlur	94.19 / **99.94**	99.28 / **100.0**	99.76 / **100.0**	99.86 / **100.0**	99.80 / **100.0**
	DirtyLens	93.37 / **99.94**	95.31 / **99.93**	95.66 / **99.96**	95.37 / **99.92**	97.43 / **99.96**
	Exposure	91.39 / **99.87**	91.00 / **99.85**	90.71 / **99.88**	90.58 / **99.85**	90.68 / **99.87**
	Snow	93.64 / **99.94**	96.50 / **99.94**	94.44 / **99.95**	94.22 / **99.95**	95.25 / **99.92**
	Haze	95.52 / **99.95**	98.35 / **99.99**	99.28 / **100.0**	99.71 / **99.99**	99.94 / **100.0**
	Decolor	93.51 / **99.96**	93.55 / **99.96**	90.30 / **99.82**	89.86 / **99.75**	90.43 / **99.83**
CURE-TSR	Noise	25.46 / **50.20**	47.54 / **63.87**	47.32 / **81.20**	66.19 / **91.16**	83.14 / **94.81**
	LensBlur	48.06 / **72.63**	71.61 / **87.58**	86.59 / **92.56**	92.19 / **93.90**	94.90 / **95.65**
	GaussianBlur	66.44 / **83.07**	77.67 / **86.94**	93.15 / **94.35**	80.78 / **94.51**	**97.36** / 96.53
	DirtyLens	29.78 / **51.21**	29.28 / **59.10**	46.60 / **82.10**	73.36 / **91.87**	98.50 / **98.70**
	Exposure	74.90 / **88.13**	**99.96** / 96.78	**99.99** / 99.26	**100.0** / 99.80	**100.0** / 99.90
	Snow	28.11 / **61.34**	61.28 / **80.52**	89.89 / **91.30**	**99.34** / 96.13	**99.98** / 97.66
	Haze	66.51 / **95.83**	97.86 / **99.50**	**100.0** / 99.95	**100.0** / 99.87	**100.0** / 99.88
	Decolor	48.37 / **62.36**	60.55 / **81.30**	71.73 / **89.93**	87.29 / **95.42**	89.68 / **96.91**

Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Gradient-based Uncertainty

Uncertainty in Detecting Challenging Conditions

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

Dataset	Method / Corruption	Mahalanobis [12] / Ours Level 1	Level 2	Level 3	Level 4	Level 5
CIFAR-10-C	Noise	96.63 / **99.95**	98.73 / **99.97**	99.46 / **99.99**	99.62 / **99.97**	99.71 / **99.99**
	LensBlur	94.22 / **99.95**	97.51 / **99.99**	99.26 / **100.0**	99.78 / **100.0**	99.89 / **100.0**
	GaussianBlur	94.19 / **99.94**	99.28 / **100.0**	99.76 / **100.0**	99.86 / **100.0**	99.80 / **100.0**
	DirtyLens	93.37 / **99.94**	95.31 / **99.93**	95.66 / **99.96**	95.37 / **99.92**	97.43 / **99.96**
	Exposure	91.39 / **99.87**	91.00 / **99.85**	90.71 / **99.88**	90.58 / **99.85**	90.68 / **99.87**
	Snow	93.64 / **99.94**	96.50 / **99.94**	94.44 / **99.95**	94.22 / **99.95**	95.25 / **99.92**
	Haze	95.52 / **99.95**	98.35 / **99.99**	99.28 / **100.0**	99.71 / **99.99**	99.94 / **100.0**
	Decolor	93.51 / **99.96**	93.55 / **99.96**	90.30 / **99.82**	89.86 / **99.75**	90.43 / **99.83**
CURE-TSR	Noise	25.46 / **50.20**	47.54 / **63.87**	47.32 / **81.20**	66.19 / **91.16**	83.14 / **94.81**
	LensBlur	48.06 / **72.63**	71.61 / **87.58**	86.59 / **92.56**	92.19 / **93.90**	94.90 / **95.65**
	GaussianBlur	66.44 / **83.07**	77.67 / **86.94**	93.15 / **94.35**	80.78 / **94.51**	**97.36** / 96.53
	DirtyLens	29.78 / **51.21**	29.28 / **59.10**	46.60 / **82.10**	73.36 / **91.87**	98.50 / **98.70**
	Exposure	74.90 / **88.13**	**99.96** / 96.78	**99.99** / 99.26	**100.0** / 99.80	**100.0** / 99.90
	Snow	28.11 / **61.34**	61.28 / **80.52**	89.89 / **91.30**	**99.34** / 96.13	**99.98** / 97.66
	Haze	66.51 / **95.83**	97.86 / **99.50**	**100.0** / 99.95	**100.0** / 99.87	**100.0** / 99.88
	Decolor	48.37 / **62.36**	60.55 / **81.30**	71.73 / **89.93**	87.29 / **95.42**	89.68 / **96.91**

Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

OLIVES @GeorgiaTech

Georgia Tech

Out-of-Distribution Detection

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

Train set →  MNIST

Goal: To detect that these datasets are not part of training

SVHN

CIFAR10

TinyImageNet

LSUN

Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Out-of-Distribution Detection

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

Dataset Distribution In	Out	Detection Accuracy	AUROC	AUPR
		Baseline [5] / ODIN [6] / Mahalanobis (V) [7] / Mahalanobis (P+FE) [7] / Ours		
CIFAR-10	SVHN	83.36 / 88.81 / 79.39 / 91.95 / **98.04**	88.30 / 94.93 / 85.03 / 97.10 / **99.84**	88.26 / 95.45 / 86.15 / 96.12 / **99.98**
	TinyImageNet	84.01 / 85.21 / 83.60 / **97.45** / 86.17	90.06 / 91.86 / 88.93 / **99.68** / 93.18	89.26 / 91.60 / 88.59 / **99.60** / 92.66
	LSUN	87.34 / 88.42 / 85.02 / **98.60** / 98.37	92.79 / 94.48 / 90.11 / **99.86** / **99.86**	92.30 / 94.22 / 89.80 / 99.82 / **99.87**
SVHN	CIFAR-10	79.98 / 80.12 / 74.10 / 88.84 / **97.90**	81.50 / 81.49 / 79.31 / 95.05 / **99.79**	81.01 / 80.95 / 80.83 / 90.25 / **98.11**
	TinyImageNet	81.70 / 81.92 / 79.35 / 96.17 / **97.74**	83.69 / 83.82 / 83.85 / 99.23 / **99.77**	82.54 / 82.60 / 85.50 / **98.17** / 97.93
	LSUN	80.96 / 81.15 / 79.52 / 97.50 / **99.04**	82.85 / 82.98 / 83.02 / 99.54 / **99.93**	81.97 / 82.01 / 84.67 / 98.84 / **99.21**

Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Out-of-Distribution Detection

Dataset Distribution		Detection Accuracy	AUROC	AUPR
In	Out	Baseline [5] / ODIN [6] / Mahalanobis (V) [7] / Mahalanobis (P+FE) [7] / Ours		
CIFAR-10	SVHN	83.36 / 88.81 / 79.39 / 91.95 / **98.04**	88.30 / 94.93 / 85.03 / 97.10 / **99.84**	88.26 / 95.45 / 86.15 / 96.12 / **99.98**
	TinyImageNet	84.01 / 85.21 / 83.60 / **97.45** / 86.17	90.06 / 91.86 / 88.93 / **99.68** / 93.18	89.26 / 91.60 / 88.59 / **99.60** / 92.66
	LSUN	87.34 / 88.42 / 85.02 / **98.60** / 98.37	92.79 / 94.48 / 90.11 / **99.86** / **99.86**	92.30 / 94.22 / 89.80 / 99.82 / **99.87**
SVHN	CIFAR-10	79.98 / 80.12 / 74.10 / 88.84 / **97.90**	81.50 / 81.49 / 79.31 / 95.05 / **99.79**	81.01 / 80.95 / 80.83 / 90.25 / **98.11**
	TinyImageNet	81.70 / 81.92 / 79.35 / 96.17 / **97.74**	83.69 / 83.82 / 83.85 / 99.23 / **99.77**	82.54 / 82.60 / 85.50 / **98.17** / 97.93
	LSUN	80.96 / 81.15 / 79.52 / 97.50 / **99.04**	82.85 / 82.98 / 83.02 / 99.54 / **99.93**	81.97 / 82.01 / 84.67 / 98.84 / **99.21**

Numbers

SVHN

CIFAR10

TinyImageNet

LSUN

Objects, natural scenes

Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Out-of-Distribution Detection

Probing the Purview of Neural Networks via Gradient Analysis

SCAN ME

Dataset Distribution		Detection Accuracy	AUROC	AUPR
In	Out	Baseline [5] / ODIN [6] / Mahalanobis (V) [7] / Mahalanobis (P+FE) [7] / Ours		
CIFAR-10	SVHN	83.36 / 88.81 / 79.39 / 91.95 / **98.04**	88.30 / 94.93 / 85.03 / 97.10 / **99.84**	88.26 / 95.45 / 86.15 / 96.12 / **99.98**
	TinyImageNet	84.01 / 85.21 / 83.60 / **97.45** / 86.17	90.06 / 91.86 / 88.93 / **99.68** / 93.18	89.26 / 91.60 / 88.59 / **99.60** / 92.66
	LSUN	87.34 / 88.42 / 85.02 / **98.60** / 98.37	92.79 / 94.48 / 90.11 / **99.86** / **99.86**	92.30 / 94.22 / 89.80 / 99.82 / **99.87**
SVHN	CIFAR-10	79.98 / 80.12 / 74.10 / 88.84 / **97.90**	81.50 / 81.49 / 79.31 / 95.05 / **99.79**	81.01 / 80.95 / 80.83 / 90.25 / **98.11**
	TinyImageNet	81.70 / 81.92 / 79.35 / 96.17 / **97.74**	83.69 / 83.82 / 83.85 / 99.23 / **99.77**	82.54 / 82.60 / 85.50 / **98.17** / 97.93
	LSUN	80.96 / 81.15 / 79.52 / 97.50 / **99.04**	82.85 / 82.98 / 83.02 / 99.54 / **99.93**	81.97 / 82.01 / 84.67 / 98.84 / **99.21**

More similar datasets (objects)

TinyImageNet

CIFAR10

LSUN

SVHN

Lee, Jinsol, et al. "Probing the Purview of Neural Networks via Gradient Analysis." *IEEE Access* 11 (2023): 32716-32732.

Objectives

Takeaways from Part III

- Part I: Gradients in Neural Networks
- Part 2: Gradients as Information
- **Part 3: Gradients as Uncertainty**
  - Defining Uncertainty in the context of Neural Networks
  - Anomaly Detection
    - GradCON: Gradient Constraints
  - Out-of-Distribution Detection
  - Adversarial Detection
  - Corruption Detection
- Part 4: Gradients as Expectancy-Mismatch
- Part 5: Conclusion and Future Directions